University of Innsbruck

Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model

Reto Stauffer, Jakob W. Messner, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis

Working Papers in Economics and Statistics

2016-07

Research Platform
eeecon
Empirical and Experimental Economics
University of Innsbruck
http://eeecon.uibk.ac.at/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
Research platform “Empirical and Experimental Economics”
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 7171
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://eeecon.uibk.ac.at/wopec/

For a list of recent papers see the backpages of this paper.

Spatio-Temporal Precipitation Climatology over Complex Terrain Using a Censored Additive Regression Model

Reto Stauffer
Universität Innsbruck

Georg J. Mayr
Universität Innsbruck

Jakob W. Messner
Universität Innsbruck

Nikolaus Umlauf
Universität Innsbruck

Achim Zeileis
Universität Innsbruck

Abstract

Flexible spatio-temporal models are widely used to create reliable and accurate estimates for precipitation climatologies. Most models are based on square root transformed monthly or annual means, where a normal distribution seems to be appropriate. This assumption becomes invalid on a daily time scale as the observations involve large fractions of zero-observations and are limited to non-negative values.

We develop a novel spatio-temporal model to estimate the full climatological distribution of precipitation on a daily time scale over complex terrain using a left-censored normal distribution. The results demonstrate that the new method is able to account for the non-normal distribution and the large fraction of zero-observations. The new climatology provides the full climatological distribution on a very high spatial and temporal resolution, and is competitive with, or even outperforms existing methods, even for arbitrary locations.

Keywords: climatology, precipitation, complex terrain, GAMLSS, censoring, daily resolution.

1. Introduction

Accurate knowledge about the climatology of precipitation is important for a wide scope of applications, such as agriculture, risk assessments, strategical project planning, water resource management, or tourism. For locations equipped with a precipitation measurement instrument, this task is straightforward. However, in most areas the observational network is too sparse to capture all local effects, and stations are mostly located at lower elevations due to environmental conditions and maintenance purposes.

To gain information about the amount or occurrence of precipitation for locations without measurements, information from an irregularly spaced observation network has to be brought to a finer (regular) region-wide grid, known as interpolation. First methods for precipitation have been published early in the last century when Thiessen (1911) pointed out that simple interpolation schemes, such as nearest neighbour, or arithmetic areal means, should not be used. Precipitation is driven by many other factors, e.g., distance to mountain ranges,

geographical position, and others (Basist, Bell, and Meentemeyer 1994). Thiessen (1911) invented an areal weighted-mean scheme which includes terrain-based properties. Although this was only a first "simple" extension, today's statistical methods still follow a similar idea.

Over the last decades, several different approaches have been developed, which can be clustered into three main groups. The first one consists of *exact interpolation schemes*, including inverse distance weighting, and various forms of kriging (e.g., Biau, Zorita, von Storch, and Wackernagel 1999; Goovaerts 2000). Another class are *regional regression models*, where for every location a (simple) regression model is adjusted from only a subset of neighbouring stations. Examples are PRISM (Precipitation-elevation Regressions on Independent Slopes Model; Daly, Neilson, and Phillips 1994; Daly, Taylor, and Gibson 1997; Daly, Gibson, Taylor, Johnson, and Pasteris 2002; Daly, Halbleib, Smith, Gibson, Doggett, Taylor, Curtis, and Pasteris 2008), and Daymet (Thornton, Running, and White 1997).

A third class of interpolation methods are *smooth spline regression models*, on which this article will focus. A common form of smooth models are generalised additive models (GAM's; Guisan, Edwards Jr., and Hastie 2002), where a response quantity is described by a set of possibly non-linear functions of covariates. Feasible functions include cyclic splines to represent annual cycles, two-dimensional splines on longitude and latitude to describe spatial distribution, altitudinal effects, and many others. Spline models have been used for long-term climatologies for different quantities, such as for annual or monthly mean temperatures or precipitation sums (e.g, Boer, de Beurs, and Hartkamp 2001; Jarvis and Stuart 2001; Vicente Serrano, Sánchez, Cuadrat *et al.* 2003; Guan, Hsu, Wey, and Tsao 2009).

Most articles have been focussing on monthly or even annual mean precipitation sums only, using a power transformation to remove skewness (Box and Cox 1964). In the literature, cubic (Stidd 1973) or square root (Hutchinson 1998b) transformations have often been suggested. The optimal power parameter may vary for different climatic zones or temporal aggregation levels. For demonstration purposes Figure 1*a1* shows monthly precipitation sums, whereas Figure 1*a2* shows the same data on a square root scale. The power transformation is able to remove most of the skewness from the observed distribution.

For a wide range of applications a finer temporal resolution is needed and - beside the mean precipitation amount - additional properties of the climatological distribution are of great interest, such as the probability of precipitation, or specific quantiles. This can be achived by either creating one specific model for each of the quantities of interest, or by modelling the full climatological distribution in one single model. An accurate estimate of the full climatological distribution requires a suitable response distribution. For daily precipitation sums, the observed and square root transformed observed distribution is shown in Figure 1*b1* & 1*b2*. Three main properties can be identified:

- (i) the distribution is highly positively skewed,
- (ii) the distribution is limited to non-negative values,
- (iii) a large fraction of all observations is exactly zero (dry days)

While the power transformation is still able to remove most of the skewness, the remaining properties stay unchanged and have to be accounted separately. As precipitation is physically limited to ≥ 0 it can be seen as left-censored (Messner, Mayr, Wilks, and Zeileis 2014).

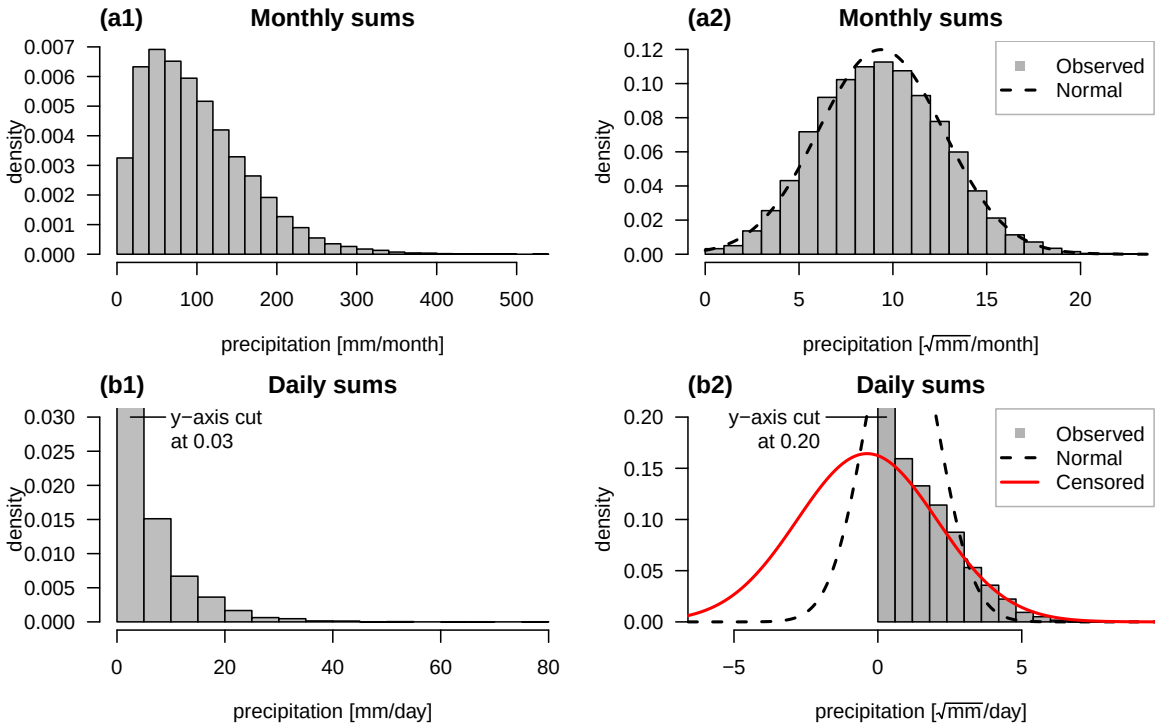

Figure 1: Density plot of precipitation sums. Top row: monthly precipitation sums from 117 stations. Bottom row: daily precipitation sums of one sample station. Right column: power-transformed observations ($\sqrt{mm}\ day^{-1}$) with a normal distribution fitted to it (black, dashed). Additinally a left-censored normal distribution is fitted to the power-transformed daily observations (bottom right; red, solid). Please note: y-axes in the bottom row are both cut.

In this article, we present a novel spatio-temporal additive model with a left-censored normal response, to estimate a full-distributional climatology of precipitation over complex terrain on a daily temporal resolution. We are using a generalised additive model for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos 2005) to create reliable estimates of the full spatio-temporal climatological distribution on a daily time scale. This allows to model the expectation, as well as the climatological variance simultaneously. To remove the (i) skewness a power transformation will be applied. To account for the two remaining properties, a left-censored normal distribution will be assumed which handles both, the (ii) lower limit at 0, and (iii) the large fraction of zero observations in the data set. The new approach allows full scalability (size of the area of interest, but also spatial- and temporal resolution) and can therefore be easily implemented and applied to new data sets.

This article is organized as follows: Section 2 introduces the concept of censoring, and the GAMLSS framework needed to estimate the high-resolution precipitation climatologies on a daily time scale. In Section 3 the area of interest and used the data set is described, followed by the climatological estimates in Section 4. While Section 4.1 shows model results, model verification and comparison will be presented in Section 4.2.

2. Methodology

2.1. Left-Censored Normal Distribution

A crucial point is the (conditional) response distribution of the model. For monthly or annual precipitation sums, a normal distribution on transformed observations using the square root works well (e.g., Hutchinson 1998a) as shown in Figure 1*a2*. The transformation is able to remove the skewness. Due to the temporal aggregation, the majority of all observations lie above zero, leading to a pseudounbounded data set (Sansom and Tait 2004), where the assumption of a normally distributed response seems appropriate. For different data sets or climatic zones, the aggregation period required to create pseudounbounded data may vary.

In contrast to the observed monthly sums, the observed distribution of daily precipitation observations is shown in Figure 1*b1* for one random station. Again, a square root transformation was applied to remove the positive skewness, shown in Figure 1*b2*. A strong peak can be seen at 0 caused by the large fraction zero observations (days without precipitation). Therefore, a left-censored normal distribution will be used in this article as a suitable response distribution. The concept of censoring is that a certain quantity cannot be observed below or above a certain threshold τ, or outside a certain range τ_1–τ_2. Precipitation is physically limited to 0 mm and can therefore be seen as left-censored at $\tau = 0$, as shown by Messner *et al.* (2014), see Figure 1*b1* & 1*b2*. In addition to the observed daily precipitation sums, Figure 1*b2* shows two fitted distributions. The dashed line shows a normal distribution based on the arithmetic mean and standard deviation, resulting in an apparently inappropriate fit. The solid line shows the fitted left-censored normal distribution. Comparing the estimated (44%) and observed (43%) probability of precipitation, as well as the estimated (2.3 $mm\ day^{-1}$) and observed (2.2 $mm\ day^{-1}$) expectation shows that the left-censored normal distribution is able to account for the large fraction of zero observations and to accurately adjust the distribution of the non-censored part. A left-censored normal distribution censored at 0 can be specified as follows:

$$y = \max(0, y^*), \quad y^* \sim \mathcal{N}(\mu, \sigma) \tag{1}$$

y^* denotes the unobservable *latent* response following a normal distribution, given the location and scale parameters μ and σ. The *observable* response y is simply the maximum of the latent response and the censoring point. From here on this distribution will be denoted as $\mathcal{N}_0$. The density (ϕ_{cens}) and the distribution function (Φ_{cens}) for $\mathcal{N}_0$ can be written as follows:

$$\phi_{cens}(x_i|\mu,\sigma,0) = \begin{cases} 0 & \text{for} \quad x_i < 0 \\ \Phi(x_i|\mu,\sigma) & \text{for} \quad x_i = 0 \\ \phi(x_i|\mu,\sigma) & \text{else} \end{cases} \tag{2}$$

$$\Phi_{cens}(x_i|\mu,\sigma,0) = \begin{cases} 0 & \text{for all:} \quad x_i < 0 \\ \Phi(x_i|\mu,\sigma) & \text{else} \end{cases} \tag{3}$$

While both quantities are set to 0 below the censoring point, both follow the density ϕ and distribution function Φ of a non-censored normal distribution, respectively, above the censoring point ($x_i > 0$). On the censoring point ($x_i = 0$) the distribution function is again equivalent to the normal distribution, while the density represents the probability that an

observation will lie exactly on 0. Therefore, the probability π to exceed 0 can be written as:

$$\pi(y > 0) = 1 - \Phi(0|\mu, \sigma) \tag{4}$$

A last property of interest is the expectation of $\mathcal{N}_0$. As the estimates will be fitted on a power-transformed scale y with $y = z^{1/p}$, this transformation has to be included to get the expectation on the original scale ($mm\ day^{-1}$). The expectation function of a power-transformed $\mathcal{N}_0$ can be expressed as (see Appendix A):

$$E[z] = \int_0^\infty z \cdot \phi(z^{\frac{1}{p}}|\mu, \sigma) \cdot \frac{z^{(\frac{1}{p}-1)}}{p} dz \tag{5}$$

Where z is already on the original scale, μ and σ are the estimated parameters of $\mathcal{N}_0$ on the power-transformed scale, and p denotes the power parameter that specifies the power-transofrmation.

2.2. Generalised Additive Model for Location, Shape, and Scale (GAMLSS)

Generalised additive models for location, scale, and shape (Rigby and Stasinopoulos 2005) are an extension to generalised additive models (GAM's; Guisan *et al.* 2002) which allow to model all parameters of a certain response distribution separately. In case of a censored normal distribution two parameters have to be specified: *latent* location (mean), and *latent* scale (standard deviation). For a left-censored normal distribution censored at $\tau = 0$ a GAMLSS model can be expressed as follows:

$$\begin{aligned} y &\sim \mathcal{N}_0(\mu, \sigma) \\ \mu &= s(\mathbf{x}) \\ \log(\sigma) &= t(\mathbf{x}) \end{aligned} \tag{6}$$

The *observable* response y is assumed to follow a left-censored normal distribution $\mathcal{N}_0$ censored at 0, with location μ and scale σ, where the log-link ensures positive values during optimization. Both distributional parameters can be expressed by a set of unknown, possibly non-linear functions $s(\dots)$ and $t(\dots)$, also known as linear predictors given the explanatory variables $\mathbf{x}$ including the covariates, such as altitude, longitude, latitude, and others.

The linear predictors can include different additive effects, such as linear effects, non-linear effects, cyclic effects, two-dimensional surfaces, and many others. Common forms of splines are e.g., thin plate splines, or B-splines (Wood 2006, Chap. 4.1, Fahrmeir, Kneib, Lang, and Marx 2013). An additional penalization allows to control the wiggliness of a spline effect leading to smooth splines, which can be defined one- or multi-dimensional to allow for complex smooth effects.

For applications where only the mean is of interest, the scale parameter in Equation 6 could be specified as a constant, leading to a homoscedastic GAM model where the variance is constant among all observations. Models of this type have been used frequently for the application of precipitation climatologies, such as in Hutchinson (1998a,b); Price, McKenney, Nalder, Hutchinson, and Kesteven (2000); Boer *et al.* (2001); Hong, Nix, Hutchinson, and Booth

(2005). However, as we would like to estimate the full daily climatological distribution, the linear predictor for $\log(\sigma)$ in Equation 6 has to be specified as well.

For the specific application of a spatio-temporal precipitation climatology, the effects $s(\mathbf{x})$ and $t(\mathbf{x})$ have to capture a possible altitudinal effect, the seasonality, as well as the spatial pattern. Therefore, the following effects have been specified for the location parameter:

$$\mu = s(\mathbf{x}) = \beta + s_1(\text{alt}) + s_2(\text{yday}) + s_3(\text{lon}, \text{lat}) + s_4(\text{yday}, \text{lon}, \text{lat}), \tag{7}$$

where β denotes the global intercept, $s_1(\text{alt})$ represents a smooth *altitudinal* effect, $s_2(\text{yday})$ a cyclic seasonal effect based on the *day of the year*, $s_3(\text{lon}, \text{lat})$ a two-dimensional spatial effect given the geographical coordinates *longitude* and *latitude*, and $s_4(\text{yday}, \text{lon}, \text{lat})$ represents a three-dimensional spline to account for spatial variabilities of the seasonal pattern across the region of interest. All smooth terms use thin-plate splines, except for the seasonal effects which use cyclic cubic splines.

Analogously to the linear predictor for the location (Equation 7), the linear predictor for the log-scale is expressed as follows:

$$\log(\sigma) = t(\mathbf{z}) = \gamma + t_1(\text{alt}) + t_2(\text{yday}) + t_3(\text{lon}, \text{lat}) + t_4(\text{yday}, \text{lon}, \text{lat}) \tag{8}$$

The two linear predictors include the same effects, as we expect the climatological variance for precipitation to also show a seasonal and spatial dependency, as well as an altitudinal effect (Equation 7 & 8). This seems appropriate for the specific task of this article, but is no general requirement for GAMLSS models.

2.3. Model Setup

The spatio-temporal precipitation climatology presented in this article is specified as in Equations 6–8. To estimate such non-parametric smooth models, including the assumption of a censored normal response, suitable software is required. We are using a novel *R* package **bamlss** (Umlauf, Zeileis, Klein, and Adler 2016b), which offers a flexible Bayesian framework for additive models for location, scale, and shape (and beyond), and the capability to handle (very) large data sets. Other frequently used software implementations to estimate smooth models are e.g., ANUSPLIN (Hutchinson 2014), or the *R* packages **mgcv** (Wood 2006), and **gamlss** (Rigby and Stasinopoulos 2005).

To compare the performance of the novel spatio-temporal climatology similar models have been estimated for each station separately using the same technique with modified linear predictors. As these climatological estimates are stationwise, only the intercepts and seasonal effects from Equation 7 & 8 have to be included ($s_2(\text{yday})$,$t_2(\text{yday})$).

To remove the skewness of the observed distribution, cubic (Stidd 1973) or square root (Hutchinson 1998b) transformations have been used in the literature. However, the power parameters used were choosen empirically and might differ on different data sets. Therefore, we were using our stationwise models to find the most suitable power parameter for this application. For each station a GAMLSS model has been fitted, optimizing the linear predictors plus a constant power parameter simultaneously. It turned out (not shown) that the optimal power parameter does not show an obvious spatial or altitudinal dependency, and varies between 1.3 and 2.0 (=square root) with median around 1.6 among all stations in the data set. Within this range, the model performance is not very sensitive to the selected power

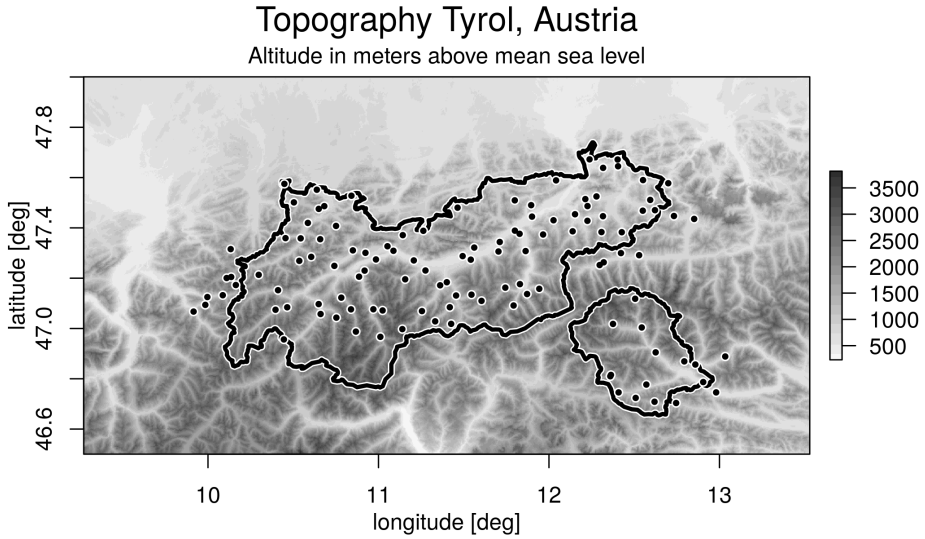

Figure 2: Topography around Tyrol, Austria. Shading indicates altitude of the topography in meters, the outline shows the border of the state of Tyrol consisting of North and East Tyrol. The black dots show the stations locations from the data set.

parameter. We therefore set the power parameter to $p = 1.6$ for all models and stations in this article.

The estimates for all GAMLSS models (*stationwise* and *spatio-temporal*; Section 4) are based on the new *R* package **bamlss**. The optimization is based on Markov-Chain Monte Carlo (MCMC) sampling in combination with an iterative weighted least squares backfitting algorithm (Umlauf, Klein, and Zeileis 2016a). Code and data used in this article can be downloaded from the **bamlss** project page (`http://bayesr.r-forge.r-project.org/`).

3. Area of Interest and Data

This article focuses on the state of Tyrol, Austria, located in Central Europe. Tyrol lies in the Eastern Alps and consists of two separated parts – North Tyrol located north of the main Alpine ridge, and East Tyrol located south of the main Alpine ridge, as shown in Figure 2. Both parts are related to the temperate climatic zone with a prevailing alpine character. The topography reaches from 465 *m amsl* up to 3798 *m amsl* including the majority of the highest mountains in Austria. This complexity is one of the main difficulties from a climatological perspective, as climatological properties can strongly vary within just a few kilometres due to topographically induced effects.

Compared to other regions, Tyrol has a relatively dense precipitation observation network. The used observation data set is provided by the local hydrographical service and includes 117 stations. Station locations are highlighted in Figure 2. Each station is equipped with a manual rain gauge to measure liquid water or liquid water equivalent accumulated over the last 24 *h*, observed at 6 UTC. The data undergo a strict quality check and correction by the maintainer. Observations are available from September 1971 through the end of 2012. 78 out of 117 stations include at least 40 years of data, 14 start within the $1980's$, 9 within

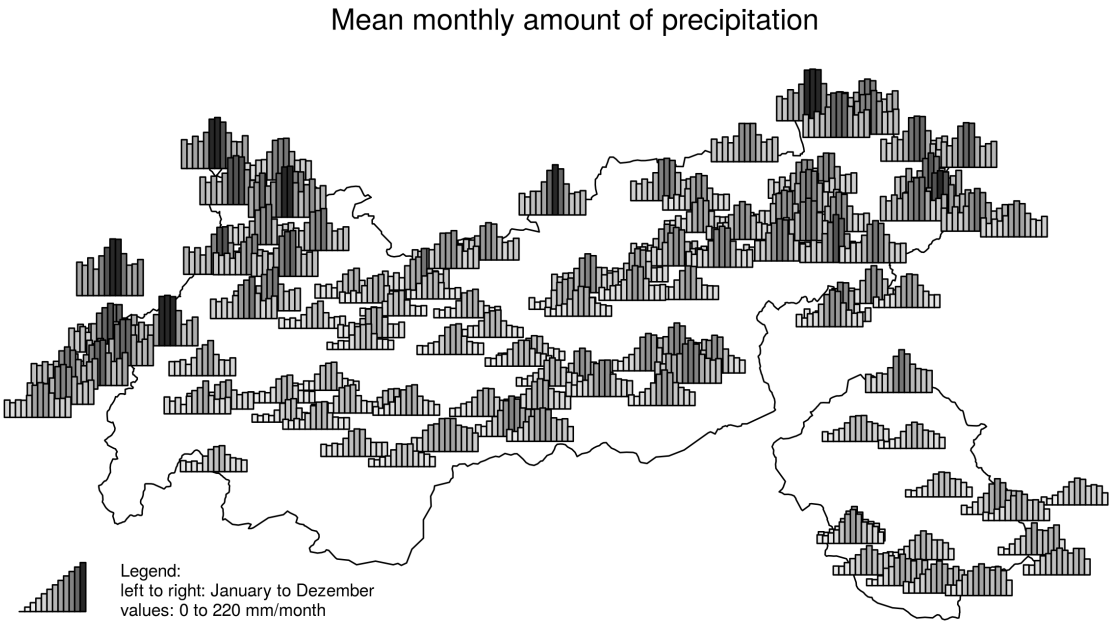

Figure 3: Mean monthly precipitation in millimetres, based on the data set. Each bar indicates one month (January–December, left to right). Bar height and luminance contain the same information.

the $1990's$ and 3 post-millennial. The total data availability is around 88% leading to a total number of roughly 1.6 million unique daily observations. The data set is freely available for non-commercial use, and can be downloaded from the **bamlss** project page (BMLFUW 2016; http://bayesr.r-forge.r-project.org/).

Figure 3 shows the mean monthly precipitation sums for all stations. The largest amounts of precipitation with around 1100–2100 mm per year are observed for the north-west and north-east stations, and a second slightly weaker maximum with more than 1000 mm per year for the south-east stations. This is due to the proximity to the foreland of the Alps (Bavaria, Germany to the north, northern Italy to the south) and dynamically driven processes. Incoming air masses get lifted when they encounter the first obstacles, leading to orographic precipitation, and a loss of moisture at the foot of the Alps (Houze 2012). On the north side, this effect is mainly caused by fronts advected from north-westerly directions, leading to higher mean precipitation amounts over the whole year. In the south-east, the highest precipitation amounts are related to mesoscale cyclones forming over the Mediterranean sea (e.g., Raulin 1879; Frei and Schär 1998). All stations show a local maximum in summer (June–August), which is mainly caused by local thermal convection, which leads to increased amounts of precipitation and thunderstorms. The convective enhancement is strongest in the pre-alpine regions north-west, and north-east of Tyrol (Wapler 2013).

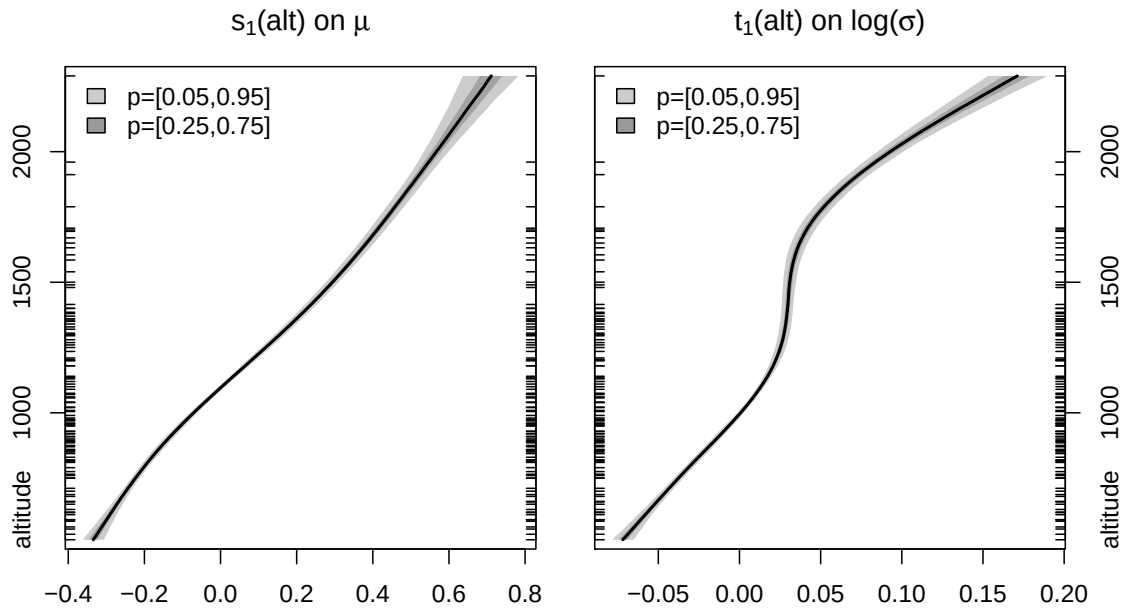

Figure 4: Centred altitudinal effects s_1(yday) on the location μ (left), and t_1(yday) on the $\log(\sigma)$ (right). Values on the power-transformed scale. Inner ticks on the ordinate indicate the altitudes of all stations in the data set. The shading shows the confidence intervals of the estimate, the width is closely related to the large amount of training data.

4. Results

First, the estimated effects of the new censored spatio-temporal precipitation climatology will be shown in Section 4.1, followed by a model comparison and validation.

4.1. Results of the New Daily-Based Spatio-Temporal Model

As described in Section 2.3 a spatio-temporal GAMLSS with a left-censored normal response is used to create the long-term precipitation climatologies (Equation 6) with the linear predictors for location and log-scale as specified in Equation 7 & 8. The individual effects of the two linear predictors are shown in Figures 4–7. All figures, except the last, show centred effects on the power transformed scale.

Figure 4 shows the altitudinal effects for location (left), and log-scale (right). As expected, the amount of precipitation and the variance increase with increasing altitude (Ekhart 1948; Frei and Schär 1998). The global cyclic seasonal effects for location (left) and log-scale (right) are shown in Figure 5. The seasonal effect shows the overall dry winter conditions from December–February (compare Figure 3) with a low variability. Overall, June–August are the months with most precipitation, with increasing variability during mid to late summer. This is related to the convective season, which has its peak between July and September. During this time period, location already decreases, while the scale nearly reaches its overall maximum. Or in other words: in autumn, the overall amount of precipitation strongly decreases (relatively dry), but the variability reaches its local annual maximum. October is the overall driest

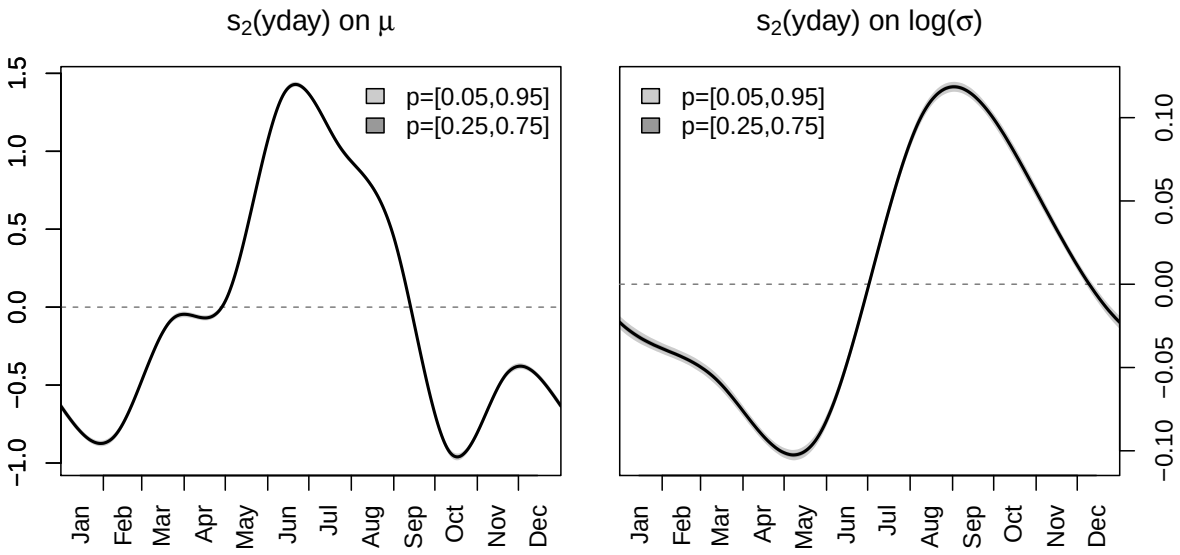

Figure 5: Centred cyclic seasonal effects s_2(yday) on location μ (left), and t_2(yday) on the log(σ) (right). Values on the power-transformed scale. The shading shows the conficence intervals of the estimate, the width is closely related to the large amount of training data. The effect controls the global seasonal effect for all stations.

month, but still shows high variability compared to the first half of the year.

The spatial effects are shown in Figure 6. As for the seasonal cycle, location and log-scale show different patterns. While location increases from south to north, the log-scale effect reaches its maximum towards the pre-alpine plains with Bavaria, Germany to the north, and Italy to the south. The increase in location is related to fronts reaching Tyrol predominantly from north and north-westerly directions. The increase in the variability is mainly caused by higher convective activity (Wapler 2013), and the orographic precipitation produced when air masses approaching from plains encounter the first higher obstacles.

Seasonal patterns differ for different regions. The three dimensional thin-plate splines s_4,t_4 in Equation 7 & 8 allow for a spatial variation of the cyclic seasonal pattern across the area of interest. This effect can be seen in Figure 7, which shows the estimated climatological expectation in $mm\ day^{-1}$ for all 117 stations in the data set. The results show that the new climatology is able to capture the different seasonal characteristics between the sub-regions north and south of the main alpine ridge.

As the new climatology returns estimates for the full distribution, one could also look at other properties such as quantiles or the probability of precipitation. Figure 8 shows the climatological distribution and the corresponding climatological estimates for two sample stations of the data set. Station A is located north of the main alpine ridge and close to the pre-alpine foreland. Station B lies south of the main alpine ridge. A few distinct features can be identified. Station A receives precipitation more frequently and observes larger amounts of precipitation. Furthermore, the different seasonalities can be seen. While station A shows a clear summer-signal with a strong increase during May–June and a corresponding decrease in autumn, station B shows a smoother transition over the whole year, with an overall lower

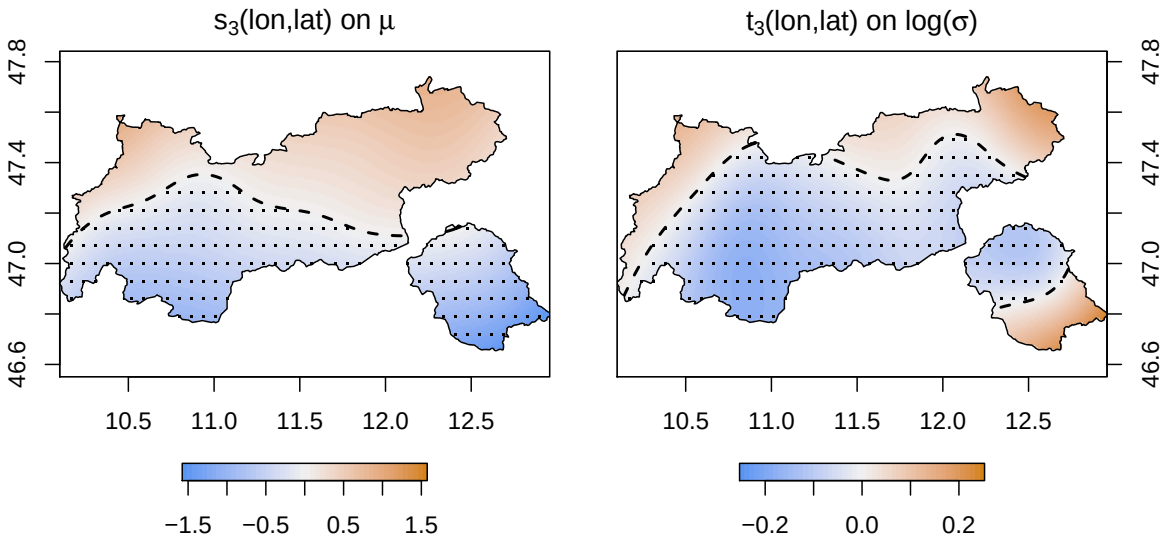

Figure 6: Centred spatial effect $s_2(\text{lon}, \text{lat})$ on the location μ (left), and $t_3(\text{lon}, \text{lat})$ on the $\log(\sigma)$ (right). Values on the power-transformed scale. Positive values orange, negative values blue and additionally dotted. The effect controls the mean underlying climatological spatial distribution of precipitation.

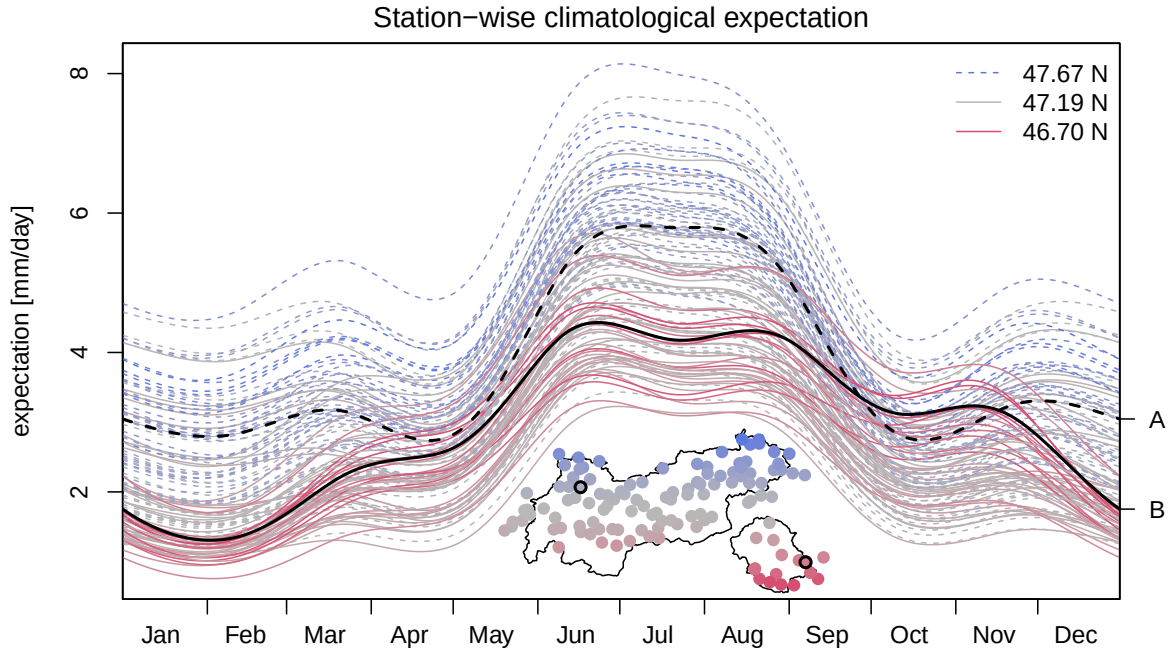

Figure 7: Expectation in $mm\ day^{-1}$ for all 117 stations used. Station coded in blue/dashed to the north, and red/solid to the south. The two sample stations A and B as shown in Figure 8 are highlighted in black. The difference in the seasonal pattern between north and south results from the tri-variate thin-plate splines s_4,t_4 based on the day of the year, longitude, and latitude.

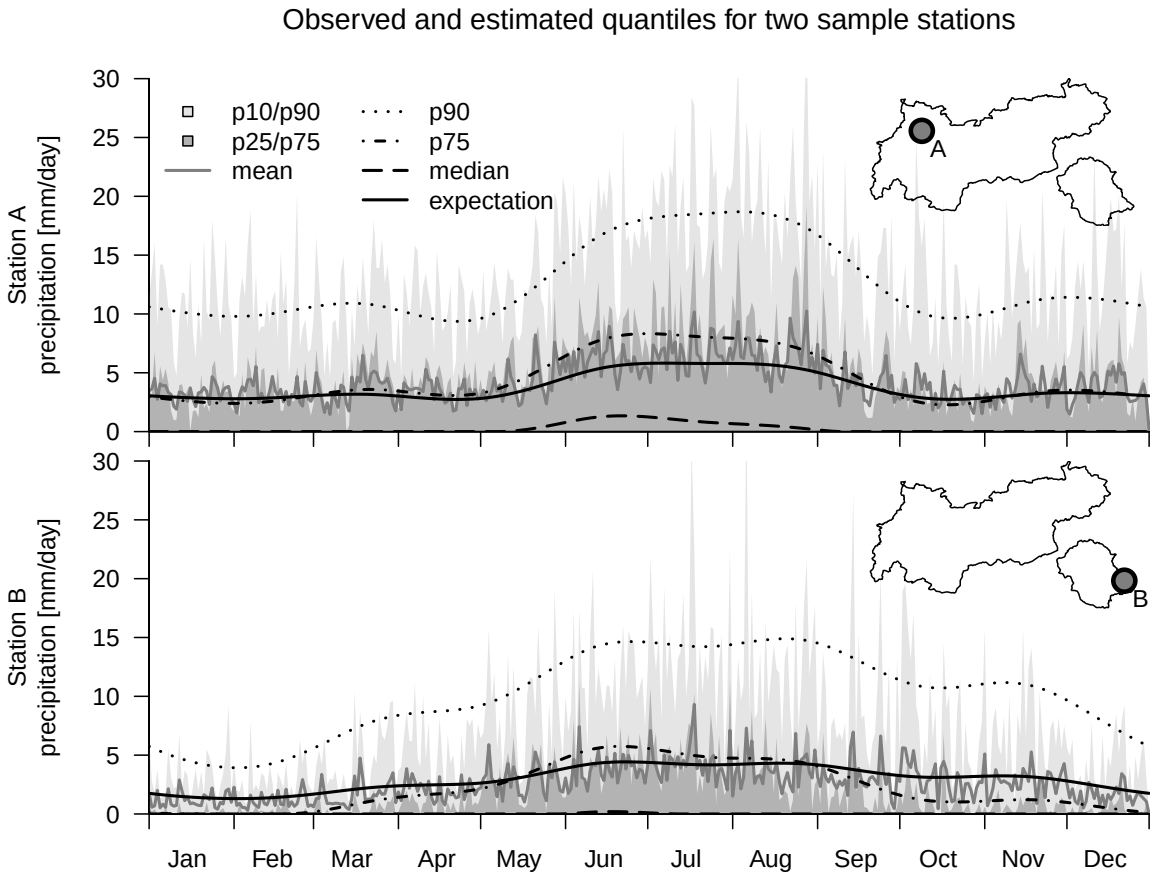

Figure 8: Distribution of daily observed precipitation sums for two sample stations. The long-term daily distribution is shown in grey including 42 years of observations: 10-90% and 25-75% inner-quantile ranges (shaded), and mean (solid, grey). In addition, the climatological estimate from the spatio-temporal model is shown. Expectation (Equation 5) as solid, and quantiles as black lines of different styles. Mean annual precipitation sums/frequency of observed precipitation for both stations: station A "Namlos" (top) 1577 mm year^{-1}/48%, station B "Iselsberg-Penzelberg" (bottom) 954 mm year^{-1}/36%.

amplitude. The censored daily spatio-temporal climatology captures the main features of amplitude, seasonality, and the overall distribution.

Figure 9 shows the spatio-temporal climatology for two sample days, January 1 (top), and the June 1 (bottom). The climatological expectation (left column) shows the overall drier winter conditions and the distinct altitudinal dependence with up to ~ 7 $mm\ day^{-1}$ on January 1, and up to ~ 10 $mm\ day^{-1}$ on June 1. The right column shows the probability of precipitation in percent. On January 1 the highest probability of observing precipitation is towards the foreland to the north, while the inner-alpine regions close to the main Alpine ridge show relatively low probabilities. On June 1 the overall probability of precipitation increases, with probabilities above $\sim 55\%$ for all mountainous areas.

4.2. Model Comparison and Validation

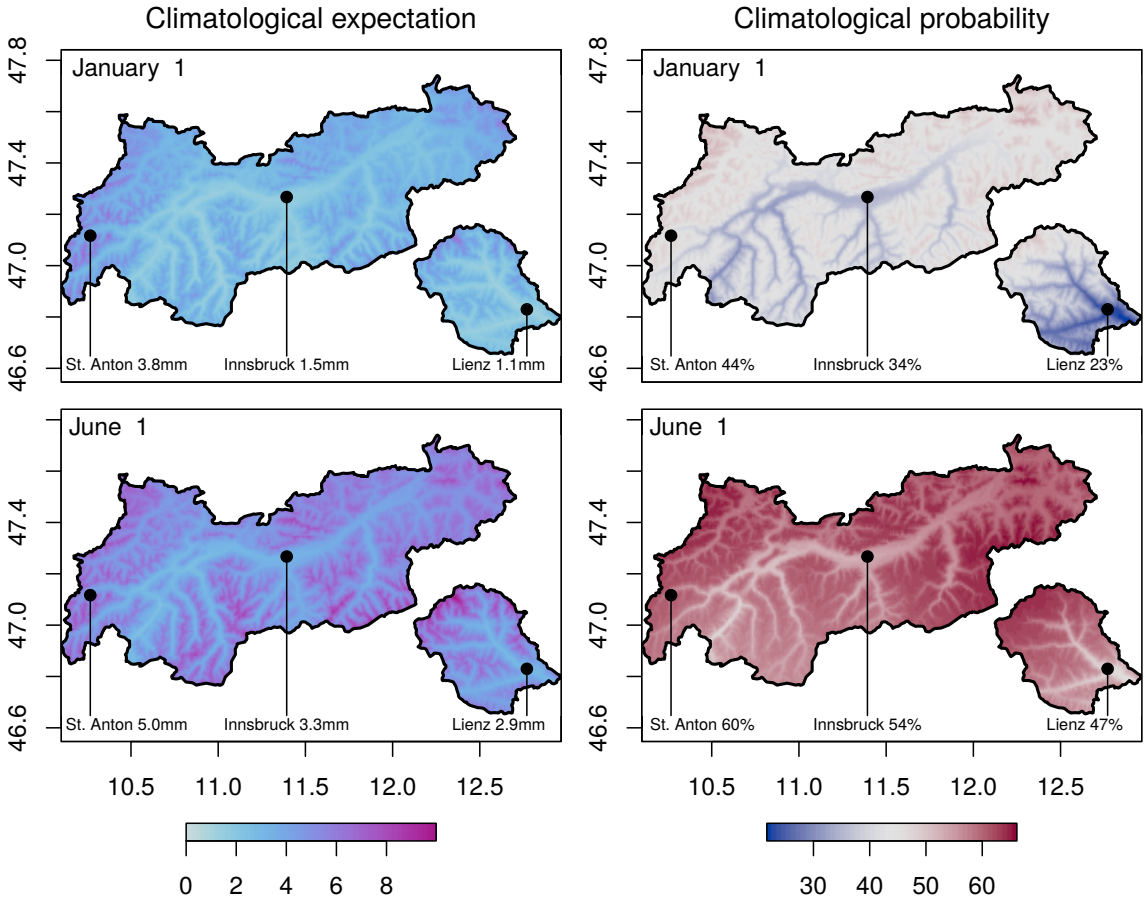

Figure 9: Climatological expectation (left; $mm\ day^{-1}$), and climatological probability of precipitation (right; %) for January 1 (top panel), and June 1 (bottom panel) respectively. Values explicitly shown for three locations: St. Anton (1284 *m amsl*), Innsbruck (574 *m amsl*), and Lienz (673 *m amsl*). Prediction based on the SRTM digital elevation model (CGIAR-CSI 2016).

The novel spatio-temporal precipitation climatology will be validated considering different aspects. Of special interest is the performance for full out-of-sample events to show the predictive performance for future (temporally out-of-sample) events at arbitrary locations within the area of interest (spatially out of sample). Three different models are estimated and are compared in this section. All models are trained on observations through the end of 2009, including up to 39 years of data (see Section 3), evaluated on the remaining three years between 2010 and 2012, from here on referred to as *training* and *test data set*.

Monthly mean model. As a robust and simple baseline reference model, long-term monthly means of the measurements are computed for each station separately. Similarly, the probability of precipitation is the long-term mean frequency of observations > 0 for a given station and month. Months with missing data are excluded.

Stationwise GAMLSS. To validate the goodness of fit of the spatial effects of the spatio-temporal model, stationwise GAMLSS climatologies with a left-censored normal distribution have been estimated. One model is estimated for each of the 117 stations using Equations 6–8 with modified linear predictors. As these models are stationwise, only the intercepts and seasonal effects have to be included.

Spatio-temporal GAMLSS. To score the predictive skill of the novel spatio-temporal climatology, a 10-fold cross validation was performed. For each cross-fold, a random subset including 10% of all stations is removed. The spatio-temporal model is estimated on the remaining stations using the specifications of Equations 6–8. For the left-out 10% of the stations the predictions are made on the remaining test data set. This leads to *spatially out-of-sample* predictions.

Measure of performance. As a measure of performance, mean absolute errors, root mean square errors, and Brier scores (Brier 1950) will be shown. While the first two are used for the amount of precipitation, the Brier scores show the performance on the estimated probability of precipitation. Mean absolute errors are based on the median of the climatological distribution ($max(0, y)^p$; Equation 1), while the root mean square errors are based on the expectation (Equation 5). The Brier scores depend on the probability that precipitation will be observed (Equation 4), with 0 as Brier score for a perfect model. To compare the different models, error-differences are shown in Figure 10. Each box-whisker is based on 117 values, each of which the mean error difference of a specific station. The error-differences are shown between each pair of methods, where the difference is defined as "method A - method B" threading "method B" as the reference. For example: Figure 10*a* shows the differences in mean absolute error (MAE), where the first pair shows "*monthly mean model* (monmean) vs. *stationwise GAMLSS* (station)". On the test data set the "*monthly mean model*" performs slightly better than the "*stationwise GAMLSS*", while both are more or less identical (in median) evaluated on the training data set. Subfigure 10*b* & 10*c* show the same validation for the root mean squared error, and Brier score respectively. The novel spatio-temporal left-censored GAMLSS model shows comparable results in all measures, or is even slightly better in terms of Brier scores.

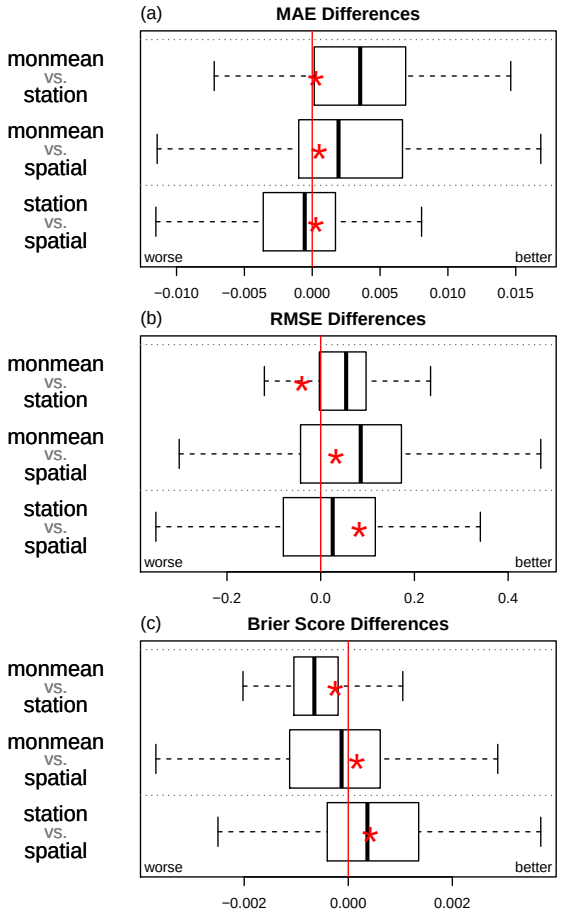

Figure 10: Differences in mean absolute error (MAE), root mean squared errors (RMSE), and Brier scores for all model pairs: monthly mean model (*monmean*), stationwise GAMLSS (*station*), and spatio-temporal GAMLSS (*spatial*). Each box-whisker consists of 117 stationwise values, each of which is the mean error for one specific station. Box-whiskers show the results on the test data set (0.25/0.5/0.75 quantiles plus additional 1.5 inner-quantile range), the red asteriks indicates the median of the same analysis on the training data set. Positive values indicate that "method A" performes better than "method B" for each "A vs. B". Absolute values lie around 3.35 (MAE), 7.25 (RMSE), and 0.24 (Brier score).

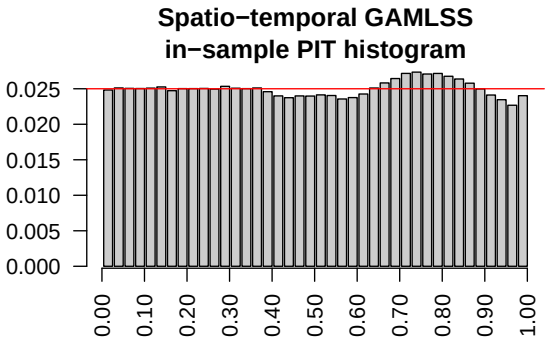

Figure 11: Pit histogram of the spatio-temporal GAMLSS model evaluated on the training data set. Width of the bins: 2.5%.

To check the suitability of the full distributional fit, a PIT histogram (Raftery, Gneiting, Balabdaoui, and Polakowski 2005) is shown in Figure 11 based on the training data set of the spatio-temporal model including all stations. The PIT histogram indicates that the left-censored normal distribution seems suitable for the application of precipitation. However, the deviation from the horizontal line indicates that there is still some room for improvement.

To sum up: the predictive skill of the novel spatio-temporal censored GAMLSS model is competitive to stationwise estimates, even for spatially out-of-sample events. This shows that the high-resolution spatio-temporal estimates generated using the method presented in this article is able to accurately reproduce the full climatological distribution of precipitation over complex terrain.

5. Conclusion and Discussion

A new method for estimating a spatio-temporal precipitation climatology with a full-distributional response, and a daily temporal resolution is presented in this article. The climatology is represented by a generalised additive model for location, scale, and shape, using the new *R* package **bamlss** (Umlauf *et al.* 2016b) for a Bayesian optimization of the regression coefficients. The estimated effects are shown in Section 4.1, and return interpretable and highly significant climatological features. An advantage of a full-distributional model is that a variety of properties can be derived from the estimate. In addition to the expectation, the probability of precipitation was verified in Section 4.2. The novel climatology with a daily temporal resolution shows a good overall performance for the amount of precipitation on the daily scale, as well as for the probability of precipitation. The results demonstrate that the concept of censoring is suitable to account for the high number of zero observations. The verification shows that the new method is competitive with, or even slightly outperforms stationwise estimates, even for arbitrary locations as summarized in Figure 10.

A PIT histogram to check the full-distributional skill is shown in Figure 11. The PIT histogram shows that the model is overall well calibrated, but there is room for improvements. Further adjustments of all tuning parameters (location, scale, but also the power parameter) might have a positive effect on the results. Beside optimizing the parameters of the left-censored normal distribution, a different response distribution might bring additional benefits. Such distributions could be e.g., a left-censored logistic distribution (Messner *et al.* 2014), or a gamma distribution. Rust, Vrac, Sultan, and Lengaigne (2013) have shown that a gamma distribution works well for precipitation on the original scale without the need to apply a power-transformation, which might distort the data. On the other hand, the gamma distribution is not defined at 0. While Scheuerer and Hamill (2015) use a censored, shifted gamma distribution, Rust *et al.* (2013) use a two-part approach where the probability of precipitation is modelled independently from the amount of precipitation. This allows to use the gamma distribution, but has the necessity to specify and estimate two different models. An advantage of the new approach presented in this article is that only one single model has to be specified to obtain probabilities, quantiles, and quantities.

A direct comparison against more complex existing methods would be needed to explicitly highlight advantages and drawbacks of our method, but needs some extensions to our current model. Adding additional covariates beside the day of the year, longitude, latitude, and altitude could further improve the model results as shown in previous publications. Conceivable covariates could be e.g., steepness and facing of the slopes, or the distance to the closest open water source. Furthermore, the new model allows to add daily covariates, such as mean wind direction, covariates explaining the regional weather situation, and many others. Some covariates have been tested but have not brought the expected results yet. Due to relatively high computational costs, estimating the full model including a "random" set of covariates will be unsatisfying. One idea would be an automated iterative variable selection approach, such as boosting or ridge-regression. Furthermore, applying the method to other censored variables would be worthwhile, such as wind speed, sunshine duration, or relative humidity.

6. Acknowledgements

Ongoing project funded by the **Austrian Science Fund (FWF)**: TRP 290-N26. The computational results presented have been achieved in part using the **Vienna Scientific Cluster (VSC)**. Data set provided by the "**Federal Ministry of Agriculture, Forestry, Environment and Water Management** (BMLFUW), Abteilung IV/4 – Wasserhaushalt (`http://ehyd.gv.at`).

Appendix A

Derivation of the expectation function for a power transformed left-censored normal distribution as in Equation 5.

Assume a left-censored normal distribution with a censoring point at 0 without a power transformation. The distribution function $F(x)$, and the density function $f(x)$ are defined as follows:

$$F(x), \quad f(x) = \frac{\partial F(x)}{\partial x}, \tag{9}$$

and therefore the expectation of the distribution becomes:

$$E\big[x\big] = \int_{x=0}^{\infty} x \cdot f(x) dx \tag{10}$$

For a left-censored normal power-transformed distribution, the distribution function $G(z)$ and density function $g(z)$ can be written in the same way, where x from Equation 9 is simply $z^{1/p}$:

$$g(z) = \frac{\partial G(z)}{\partial z} = \frac{\partial F(z^{1/p})}{\partial z}, \tag{11}$$

and therefore:

$$\frac{\partial F(z^{1/p})}{\partial z} = f(z^{1/p}) \cdot \frac{z^{(\frac{1}{p}-1)}}{p}, \tag{12}$$

leading to Equation 5 as shown in the article.

References

Basist A, Bell GD, Meentemeyer V (1994). "Statistical Relationships between Topography and Precipitation Patterns." *Journal of Climate*, **7**(9), 1305–1315. doi:10.1175/1520-0442(1994)007<1305:SRBTAP>2.0.CO;2.

Biau G, Zorita E, von Storch H, Wackernagel H (1999). "Estimation of Precipitation by Kriging in the EOF Space of the Sea Level Pressure Field." *Journal of Climate*, **12**(4), 1070–1085. doi:10.1175/1520-0442(1999)012<1070:EOPBKI>2.0.CO;2.

BMLFUW (2016). "Bundesministerium für Land und Forstwirtschaft, Umwelt und Wasserwirtschaft (BMLFUW), Abteilung IV/4 - Wasserhaushalt." http://ehyd.gv.at. Accessed: 2016-02-29.

Boer EP, de Beurs KM, Hartkamp AD (2001). "Kriging and Thin Plate Splines for Mapping Climate Variables." *International Journal of Applied Earth Observation and Geoinformation*, **3**(2), 146–154. doi:10.1016/S0303-2434(01)85006-6.

Box GE, Cox DR (1964). "An analysis of transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 211–252.

Brier GW (1950). "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, **78**(1), 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

CGIAR-CSI (2016). "SRTM 90m Digital Elevation Database v4.1." http://srtm.csi.cgiar.org. Accessed: 2016-02-29.

Daly C, Gibson WP, Taylor GH, Johnson GL, Pasteris P (2002). "A Knowledge-Based Approach to the Statistical Mapping of Climate." *Climate Research*, **22**(2), 99–113. doi:10.3354/cr022099.

Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP (2008). "Physiographically Sensitive Mapping of Climatological Temperature and Precipitation Across the Conterminous United States." *International Journal of Climatology*, **28**(15), 2031–2064. doi:10.1002/joc.1688.

Daly C, Neilson RP, Phillips DL (1994). "A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain." *Journal of Applied Meteorology*, **33**(2), 140–158. doi:10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Daly C, Taylor G, Gibson W (1997). "The PRISM Approach to Mapping Precipitation and Temperature." *Proceeding: 10th AMS Conference on Applied Climatology, Reno, NV*, pp. 208–209.

Ekhart E (1948). "Die Niederschlagsverteilung in den Alpen nach dem Anomalienprinzip." *Geografiska Annaler*, **30**, 728–739. doi:10.2307/519914.

Fahrmeir L, Kneib T, Lang S, Marx B (2013). *Regression – Models, Methods and Applications*. Springer-Verlag, Berlin. ISBN 978-3-642-34332-2.

Frei C, Schär C (1998). "A Precipitation Climatology of the Alps from High-Resolution Rain-Gauge Observations." *International Journal of Climatology*, **18**, 873–900. doi:10.1002/(SICI)1097-0088(19980630)18:8<873::AID-JOC255>3.0.CO;2-9.

Goovaerts P (2000). "Geostatistical Approaches for Incorporating Elevation into the Spatial Interpolation of Rainfall." *Journal of Hydrology*, **228**(1–2), 113–129. doi:10.1016/S0022-1694(00)00144-X.

Guan BT, Hsu HW, Wey TH, Tsao LS (2009). "Modeling Monthly Mean Temperatures for the Mountain Regions of Taiwan by Generalized Additive Models." *Agricultural and Forest Meteorology*, **149**(2), 281–290. doi:10.1016/j.agrformet.2008.08.010.

Guisan A, Edwards Jr TC, Hastie T (2002). "Generalized Linear and Generalized Additive Models in Studies of Species Distributions: Setting the Scene." *Ecological Modelling*, **157**(2–3), 89–100. doi:10.1016/S0304-3800(02)00204-1.

Hong Y, Nix HA, Hutchinson MF, Booth TH (2005). "Spatial Interpolation of Monthly Mean Climate Data for China." *International Journal of Climatology*, **25**(10), 1369–1379. doi:10.1002/joc.1187.

Houze R (2012). "Orographic effects on precipitating clouds." *Reviews of Geophysics*, **50**(1). doi:10.1029/2011RG000365.

Hutchinson M (2014). "ANUSPLIN Version 4.4. Centre for Resource and Environmental Studies." http://fennerschool.anu.edu.au/research/products/.

Hutchinson MF (1998a). "Interpolation of Rainfall Data with Thin Plate Smoothing Splines – Part I: Two Dimensional Smoothing of Data with Short Range Correlation." *Journal of Geographic Information and Decision Analysis*, **2**, 168–185.

Hutchinson MF (1998b). "Interpolation of Rainfall Data with Thin Plate Smoothing Splines – Part II: Analysis of Topographic Dependence." *Journal of Geographic Information and Decision Analysis*, **2**(2), 152–167.

Jarvis CH, Stuart N (2001). "A Comparison among Strategies for Interpolating Maximum and Minimum Daily Air Temperatures. Part II: The Interaction between Number of Guiding Variables and the Type of Interpolation Method." *Journal of Applied Meteorology*, **40**(6), 1075–1084. doi:10.1175/1520-0450(2001)040<1075:ACASFI>2.0.CO;2.

Messner JW, Mayr GJ, Wilks DS, Zeileis A (2014). "Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored." *Monthly Weather Review*, **142**(8), 3003–3014. doi:10.1175/MWR-D-13-00355.1.

Price DT, McKenney DW, Nalder IA, Hutchinson MF, Kesteven JL (2000). "A Comparison of Two Statistical Methods for Spatial Interpolation of Canadian Monthly Mean Climate Data." *Agricultural and Forest Meteorology*, **101**(2–3), 81–94. doi:10.1016/S0168-1923(99)00169-0.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005). "Using Bayesian Model Averaging to Calibrate Forecast Ensembles." *Monthly Weather Review*, **133**(5), 1155–1174. doi:10.1175/MWR2906.1.

Raulin V (1879). "Über die Verteilung des Regens im Alpengebiet von Wien bis Marseille." *Zeitschrift der österreichischen Gesellschaft für Meteorologie*, **14**, 233–247.

Rigby RA, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape, (with discussion)." *Applied Statistics*, **54**, 507–554. doi:10.1111/j.1467-9876.2005.00510.x.

Rust HW, Vrac M, Sultan B, Lengaigne M (2013). "Mapping Weather–Type Influence on Senegal Precipitation Based on a Spatial–Temporal Statistical Model." *Journal of Climate*, **26**(20), 8189–8209. doi:10.1175/JCLI-D-12-00302.1.

Sansom J, Tait A (2004). "Estimation of Long-Term Climate Information at Locations with Short-Term Data Records." *Journal of Applied Meteorology*, **43**(6), 915–923. doi:10.1175/1520-0450(2004)043$<$0915:EOLCIA$>$2.0.CO;2.

Scheuerer M, Hamill TM (2015). "Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions." *Monthly Weather Review*, **143**(11), 4578–4596. doi:10.1175/MWR-D-15-0061.1.

Stidd CK (1973). "Estimating the Precipitation Climate." *Water Resources Research*, **9**(5), 1235–1241. doi:10.1029/WR009i005p01235.

Thiessen AH (1911). "Precipitation Averages for Large Areas." *Monthly Weather Review*, **39**(7), 1082–1089. doi:10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2.

Thornton PE, Running SW, White MA (1997). "Generating Surfaces of Daily Meteorological Variables over Large Regions of Complex Terrain." *Journal of Hydrology*, **190**(3–4), 214–251. doi:10.1016/S0022-1694(96)03128-9.

Umlauf N, Klein N, Zeileis A (2016a). "bamlss: Bayesian Additive Models for Location, Scale and Shape (and Beyond)." *Unpublished manuscript*.

Umlauf N, Zeileis A, Klein N, Adler D (2016b). "bamlss: Bayesian Additive Models for Location Scale and Shape (and Beyond)." https://R-Forge.R-project.org/R/?group_id=865.

Vicente Serrano SM, Sánchez S, Cuadrat JM, *et al.* (2003). "Comparative Analysis of Interpolation Methods in the Middle Ebro Valley (Spain): Application to Annual Precipitation and Temperature." *Climate Research*, **24**(2), 161–180. doi:10.3354/cr024161.

Wapler K (2013). "High-Resolution Climatology of Lightning Characteristics within Central Europe." *Meteorology and Atmospheric Physics*, **122**(3–4), 175–184. doi:10.1007/s00703-013-0285-1.

Wood S (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Affiliation:

Reto Stauffer, Jakob W. Messner
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstraße 15
6020 Innsbruck, Austria, *and*
Institute of Atmospheric and Cryospheric Sciences
Faculty of Geo- and Atmospheric Sciences
Universität Innsbruck
Innrain 52
6020 Innsbruck, Austria
E-mail: Reto.Stauffer@uibk.ac.at, Jakob.Messner@uibk.ac.at

Nikolaus Umlauf, Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstraße 15
6020 Innsbruck, Austria
E-mail: Nikolaus.Umlauf@uibk.ac.at, Achim.Zeileis@uibk.ac.at

Georg J. Mayr
Institute of Atmospheric and Cryospheric Sciences
Faculty of Geo- and Atmospheric Sciences
Universität Innsbruck
Innrain 52
6020 Innsbruck, Austria
E-mail: Georg.Mayr@uibk.ac.at

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://eeecon.uibk.ac.at/wopec/

2016-07 **Reto Stauffer, Jakob W. Messner, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model

2016-06 **Michael Razen, Jürgen Huber, Michael Kirchler:** Cash inflow and trading horizon in asset markets

2016-05 **Ting Wang, Carolin Strobl, Achim Zeileis, Edgar C. Merkle:** Score-based tests of differential item functioning in the two-parameter model

2016-04 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Non-homogeneous boosting for predictor selection in ensemble post-processing

2016-03 **Dietmar Fehr, Matthias Sutter:** Gossip and the efficiency of interactions

2016-02 **Michael Kirchler, Florian Lindner, Utz Weitzel:** Rankings and risk-taking in the finance industry

2016-01 **Sibylle Puntscher, Janette Walde, Gottfried Tappeiner:** Do methodical traps lead to wrong development strategies for welfare? A multilevel approach considering heterogeneity across industrialized and developing countries

2015-16 **Niall Flynn, Christopher Kah, Rudolf Kerschbamer:** Vickrey Auction vs BDM: Difference in bidding behaviour and the impact of other-regarding motives

2015-15 **Christopher Kah, Markus Walzl:** Stochastic stability in a learning dynamic with best response to noisy play

2015-14 **Matthias Siller, Christoph Hauser, Janette Walde, Gottfried Tappeiner:** Measuring regional innovation in one dimension: More lost than gained?

2015-13 **Christoph Hauser, Gottfried Tappeiner, Janette Walde:** The roots of regional trust

2015-12 **Christoph Hauser:** Effects of employee social capital on wage satisfaction, job satisfaction and organizational commitment

2015-11 **Thomas Stöckl:** Dishonest or professional behavior? Can we tell? A comment on: Cohn et al. 2014, Nature 516, 86-89, "Business culture and dishonesty in the banking industry"

2015-10 **Marjolein Fokkema, Niels Smits, Achim Zeileis, Torsten Hothorn, Henk Kelderman:** Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees

2015-09 **Martin Halla, Gerald Pruckner, Thomas Schober:** The cost-effectiveness of developmental screenings: Evidence from a nationwide programme

2015-08 **Lorenz B. Fischer, Michael Pfaffermayr:** The more the merrier? Migration and convergence among European regions

2015-07 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Cooperation and discrimination within and across language borders: Evidence from children in a bilingual city

2015-07 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Cooperation and discrimination within and across language borders: Evidence from children in a bilingual city *forthcoming in European Economic Review*

2015-06 **Martin Geiger, Wolfgang Luhan, Johann Scharler:** When do Fiscal Consolidations Lead to Consumption Booms? Lessons from a Laboratory Experiment

2015-05 **Alice Sanwald, Engelbert Theurl:** Out-of-pocket payments in the Austrian healthcare system - a distributional analysis

2015-04 **Rudolf Kerschbamer, Matthias Sutter, Uwe Dulleck:** How social preferences shape incentives in (experimental) markets for credence goods *forthcoming in Economic Journal*

2015-03 **Kenneth Harttgen, Stefan Lang, Judith Santer:** Multilevel modelling of child mortality in Africa

2015-02 **Helene Roth, Stefan Lang, Helga Wagner:** Random intercept selection in structured additive regression models

2015-01 **Alice Sanwald, Engelbert Theurl:** Out-of-pocket expenditures for pharmaceuticals: Lessons from the Austrian household budget survey

University of Innsbruck

Working Papers in Economics and Statistics

2016-07

Reto Stauffer, Jakob W. Messner, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis

Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model

Abstract
Flexible spatio-temporal models are widely used to create reliable and accurate estimates for precipitation climatologies. Most models are based on square root transformed monthly or annual means, where a normal distribution seems to be appropriate. This assumption becomes invalid on a daily time scale as the observations involve large fractions of zero-observations and are limited to non-negative values. We develop a novel spatio-temporal model to estimate the full climatological distribution of precipitation on a daily time scale over complex terrain using a left-censored normal distribution. The results demonstrate that the new method is able to account for the non-normal distribution and the large fraction of zero-observations. The new climatology provides the full climatological distribution on a very high spatial and temporal resolution, and is competitive with, or even outperforms existing methods, even for arbitrary locations.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)